Research Paper | Engineering | E-ISSN No : 2454-9916 | Volume : 3 | Issue : 5 | May 2017

SOCIAL MEDIA E-CUSTOMERS' BEHAVIOUR MINING

Ms. Rekha K Dimke [1] | Mr. Aniruddh Fataniya [2]

[1] PG Student, Dept. of Comp. Engineering, PIET, Parul Universit, Waghodia, India.

[2] Assistant Professor, Dept. of Comp. Sci & Engineering, PIET, Parul University, Waghodia, India.

ABSTRACT

As nowadays, Social media has become an important source of information, where people can express their opinions and their views. To analyze social media data of e-customer's there is need to extract information for making predictions of how they behaves during shopping in an e-commerce market place. After collecting data from an e-commerce market, performed a data mining application for extracting about how customers' behaves online whether to buy product or not. The model which is presented predicts whether customers will not or will buy their items or products added to shopping baskets on a market place. As there is massive growth of online social networks (OSN) like Twitter, Facebook and other social networking portals have created a need to determine people's opinion and moods. Posting user feedback on products has become increasingly popular for people to express their opinions toward products and services. The companies think that there is a chance for an improvement in market for a product to people to aware and feel about it. In this study, there is use of sentiment dictionary as an Affine for tokenization and preprocessing process and also there is use of machine learning techniques to find about e-commerce site that is more useful and good for e-customers by predicting and analyzing through their reviews. "

KEYWORDS: Twitter Data, Sentiment dictionary, Social media, Support Vector Machine, Artificial Neural Network.

I. INTRODUCTION

E-commerce means to sell and to buy goods and services, or transmission of data or funds, through an electronic network, such as an Internet. Social media is a promising link which helps to build connection on social networks, personal information channels and mass information. Nowadays, Social media plays an important and precious role for information, where people can express their views.

When these opinions are related to company, accurate analysis can provide them with information such as quality of products and hinders that affects other customer decisions, feedback which are given earlier for launching products, trends, news of companies and also knowledge about other company which are in competition. The massive growth of online social networks (OSN) like Twitter, Facebook and other social networking portals have created a need to determine people's opinion and moods

"The advantage of an electronic market place is to offer many choices, low price, easy way to search to access customers online. Thus Internet market share is important each and every day [2]. Opinion Mining is also known as Sentiment Analysis [1]. The company think that there is a way for improvement of product in market if they are aware of how people feel about specific product. Twitter data is use for studying, analyzing data. Using data mining methods such as Support Vector Machine and ANN it can examine about customer behaviour. Classification is an important in data mining. Data mining is an area which is included in machine learning, "

With the rapid expansion of E-Commerce and social networking sites, so there is huge amount of information available in social media. Views of various products which are expressed in OSN's plays an important role in market business analysis. In this paper, using machine learning technique with Semantic analysis which is used to classify the people's opinions and views based on twitter data. Support Vector Machine is used with unigram model and Negation model for giving better performance than using it alone. Emoticons and affine dictionary is used so that it will be decided which E-commerce site is more useful and good for customer based on reviews.

II. METHODOLOGY

A. Dictionaries:

There are various mainly three types of dictionaries used in this research work as follows:

I. Afinn-111:

AFINN-111: This is new version with 2477 words and phrases.

ii. Lexicon dictionary:

Lexicon Dictionary contains 3382 words in dictionary. Its score is between -5 and +5 as affine dictionary.

iii. Hybrid dictionary:

Hybrid dictionary is combination of both lexicon and affine dictionary. It includes 3478 words and their related score.

iv. Emoticons:

An Emoticon, for example, such that ;-), will be shorthand for An facial outflow. It permits those creator should express her/his feelings, moods and emotions, What's more augments an composed message with non-verbal components. It aides on draw the reader's attention, that's more enhances Furthermore enhances the Comprehension of the meaning.

ICON	*MEANING*
:-) =) :) 8) :] =| => 8-) :-> :-] :") :')	smiley
:3 :> :') :-3 => :-> :-V =v :-I	happy face
^_^ ^L^ ^)^	happiness
:* :-*	kiss, couple kissing
;-) ;) ;-] ;] ;-> ;> %-}	wink, smirk
<3	heart
:-D :D =D :-P =3 xD	laughing
:P =P	tongue sticking out, playful
O.o o.O	surprised
:v	gape
B) B-) B| 8|	feel cool
:'-)	tears of happiness
:!	exclamation
:-X	Sealed lips, wearing braces
=* ;-* ;*	kiss

Figure 2.4: Emoticons icon and meaning

B. Classification Method

i. Support Vector Machine:

That principle guideline of SVM will be with figure out straight separators in the scan space which might best differentiate the distinctive classes. As below figure shows there would 2 classes x, o and there need aid 3 Hyper-planes A, b Also c. Hyper-plane An gives those best division between those classes, a result the ordinary separation of any of the information focuses is those largest, something like that it speaks to the greatest edge for division. Those help vectors would the information focuses that need aid closest of the dividing hyper plane; these focuses need aid on the limit of the piece. Quick information need aid ideally suiting to SVM order due to the meager nature of text, to which couple features are irrelevant, Anyhow they tend with be associated with each other Furthermore by sorted out under linearly distinct Classes.

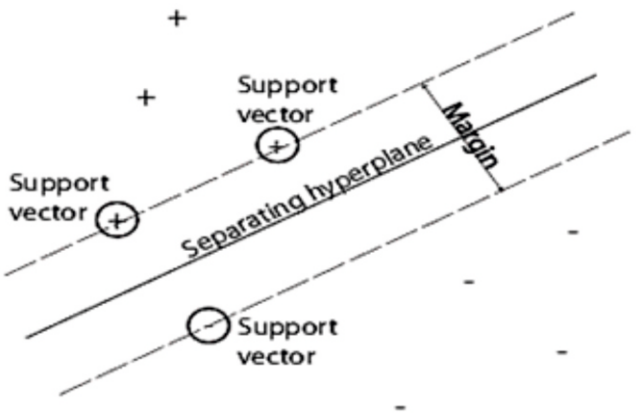


Figure 2.5.1: Support Vector Machine [12]

1) + indicating data points of type 1, and

2) - indicating data points of type –1

SVM has been used successfully in many real-world problems
1) Text categorization

2) Image classification

3) Hand-written character recognition

"Text data are ideally suited for SVM classification because of the sparse nature of text, in which few features are irrelevant, but they tend to be correlated with one another and generally organized into linearly separable categories. "

ii. ANN [3][11]
An artificial neural network consists of different artificial neurons that are linked together according to specific network architecture. A neural network has emerged as an important tool for classification. During past decade neural network classification has established as a promising alternative to various conventional classification methods. The neural network with appropriate network structure can handle the correlation/dependence between input variables.

1) The objective of the neural network is to transform the all inputs into meaningful or understandable outputs.

2) An artificial neural network (ANN), is also known as Neural Network (NN), is a computational or mathematical model which is inspired by the functional and/or structure of biological neural networks.

3) Neural Networks have performed successfully where other methods have not, predicting system behavior, recognizing and matching complicated, vague, or incomplete data patterns.

4) Application: Ann is used pattern recognition, interpretation, prediction, diagnosis.

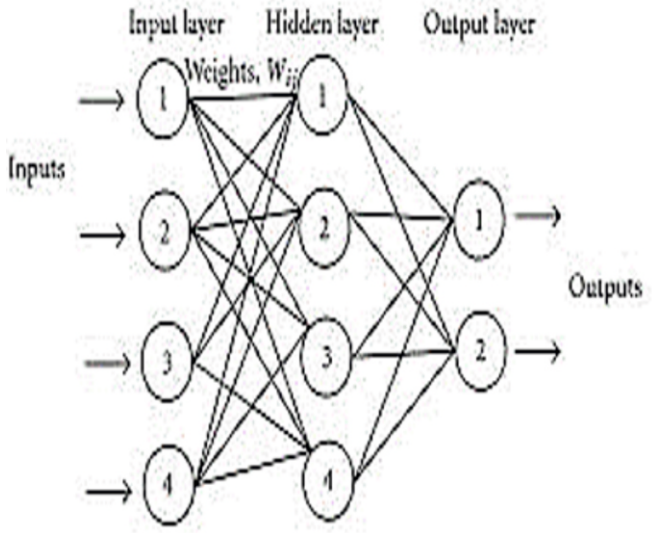
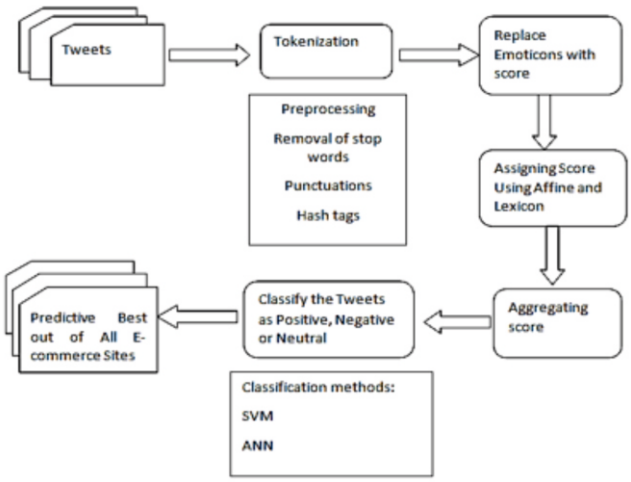


Figure 2.5.2.: Artificial Neural Network [12]

It involves intelligent opinion mining .It also contains a method which allows classifying opinion and sentiment score classes. The hybrid neural network contain of a modified probabilistic, neural network combine with a single layer classifier. The inputs of the network comprise binary images of potential opinion samplers. The consecutive bits represent evaluative words which were scored on a scale of intensity in an appropriate manner. The output provides the detection of potential opinions and the classification of opinion type classes [11].The method uses modified hybrid multilayer neural networks to recognize whether it is opinion type for finding its score. The network is a pattern classifier.

III. PROPOSED FLOW



A. Collection of tweets
"The input to the emotion analyzer is a user entered keyword based on which recent tweets which fetched from Twitter using its Search API. The Twitter Search API is a dedicated API for running searches against the real-time index of recent tweets. Each request will return up to more than 100 tweets, for a single query. By running the search script we can keep up with most search topics without missing any tweets. For our system, by gathering our dataset by consulting the Twitter API and making use of word spotting based on occurrence of the word and are querying the recent tweets."

B. Pre-Processing
"Twitter data is unstructured data. It needs to be processed before it can be used. Hence the tweets obtained are cleaned to remove unwanted discrepancies and retain only information that will help in determining the underlying emotion. This makes data easier to process in the later stages."

"The procedure for pre-processing consists of the following steps: "
1) "Removing all non-English Tweets."

2) "Converting all the tweets collected to the lower case."

3) "Removing the URLs – erased all string that describes links or hyperlinks present in the tweets."

4) "Replacing any usernames present in the tweets to @username – removed the username and because these are not considers for sentiments."

5) "Converting the hash tags to normal words because hash tags can provide some helpful information, so it is useful to replace them with the literally same word without the hash. E.g. #Happy replaced with Happy."

6) "Removing any unnecessary characters, extra spaces etc."

7) "Remove all the number from tweets and also remove all words which don't start with an alphabet, for example 9th, 9:15am."

8) "Removing punctuation like commas, single/double quotes question marks, etc. at the beginning and end of each word in a tweet. E.g. Happy!!!!!! Replaced with Happy. Replacing two or more repeating letters in a tweet by two letters of the same in tweets, sometimes users repeat letters to stress the emotion or feelings. E.g. Happy, Happyyyyyyyy for 'Happy'. "

C. Feature Extraction
"Extraction of features plays a very important concept that is responsible for the accuracy of the system. To decide what are the features that are relevant to the classifier, need a feature extractor. The one that have used returns a dictionary indicating what type of words are contained in the input. Here, input is pre-processed tweet that is first filtered using the steps which are mentioned below:"

1) "Polarity Score of the Tweet."

2) "Remove all stop words like a, the, is, etc. which don't indicate any emotion."

3) **"Replace the emotions with similar mining of word i.e. -): with happy."**

"Use of Unigram Model: The feature extraction method, extracts the aspect (adjective) from the dataset. Later this adjective is used to show the positive and negative polarity in a sentence which is useful for determining the opinion of the individuals using unigram model. Unigram model extracts the adjective and segregates it. It discards the preceding and successive word occurring with the adjective in the sentences. For above example, i.e. "Driving Happy" through unigram model, only Happy is extracted from the sentence. Once the tweets are filtered, the output of the feature extractor is a list of the feature words present in the tweet. "

D. Machine Learning Classification

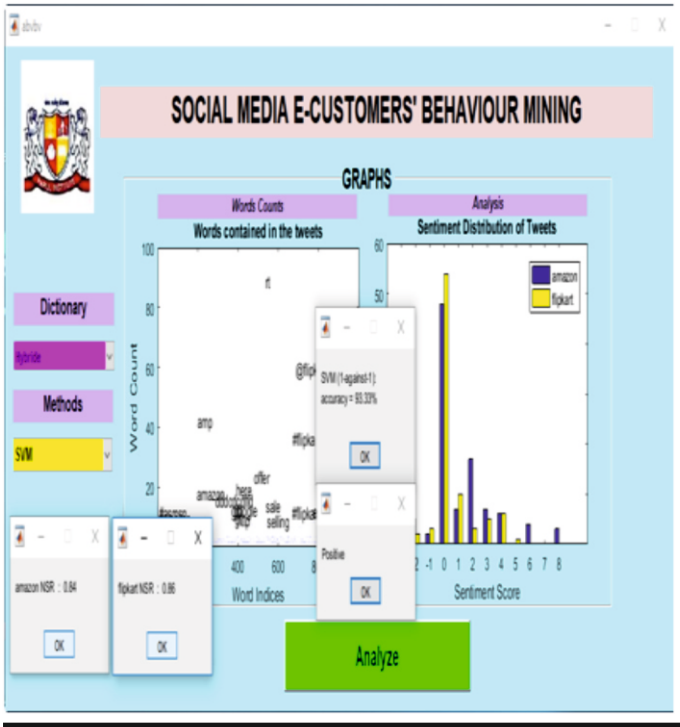
"A classifier is a learning model that associates which learning algorithms that can analyze data and recognize patterns which can be used for classification. Machine Learning Classification: The ability of a machine to improve its document classification performance based on previous results of document classification."

"Support Vector Machine: In Machine Learning Approach (MLA), SVM are supervised learning models with associated learning algorithms that analyses data for classification. If there is given a set of training examples each marked for belonging to one of the classes or categories. SVM constructs hyper plane which can be used for classification of the data. SVM is used in text categorization and it also has ability to learn .SVM uses over-fitting protection so that they have ability to handle large number of features. SVM is also use to find out polarity of textual comments. Among the different variants of SVM, the multiclass SVM is very useful for sentiment analysis. The classification algorithm for centroid first is need to calculate the centroid vector for training class. Then to find similarities between document and all the centroids that are calculated and the document is defined for a class based on these identical values. "

"Artificial Neural Network: Neural Network is an important for classification. It is network structure can handle the correlation between input variables and neural network can adjust themselves to data. A neural network plays an important tool for classification. During past decade Neural Network (NN) classification has established as a promising alternative to various conventional classification methods. The neural network with appropriate network structure can be able to handle the correlation/dependence between input variables. The advantages of neural networks are as follows: First, neural networks are data driven and are method which is self-adaptive in which they can adjust themselves to the data without any explicit specification of functional or distributional or functional form for the model. Second, they are universal functional approximates in which neural networks can approximate any function with arbitrary accuracy."

IV. EXPERIMENTAL RESULTS

A. NSR of two E-Commerce site using hybrid dictionary and SVM Classifier



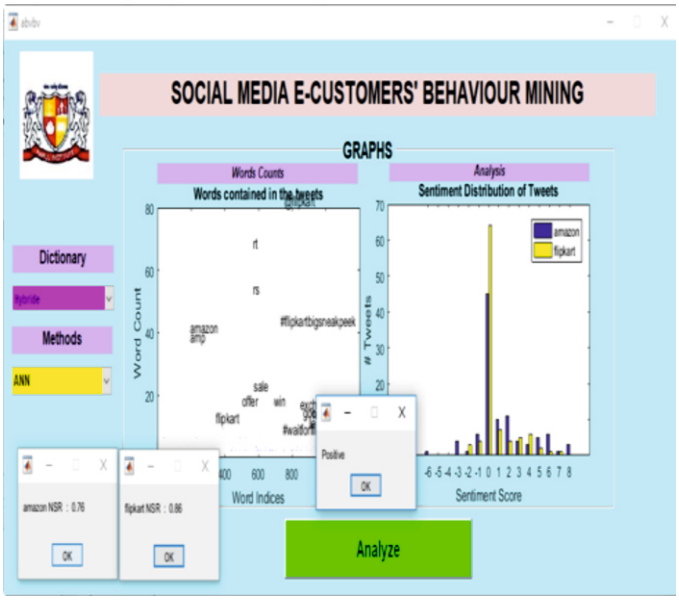
In this above screenshot, it is shown that for two combine E-commerce site i.e amazon, flip kart is selected and selecting Hybrid dictionary and SVM classifier method the above figure shows that amazon NSR is 0.86 and time taken is 3.21 sec and Svm accuracy is 93.33%. Thus amazon, flip kart belongs to positive class but thus words counts and analysis has been shown in above figure using mat lab.

B. SVM showing confusion matrix

```
SVM (1-against-1):
accuracy = 93.33%
Confusion Matrix in Persentage:
   100     0     0
     0    80    20
     0     0   100


precision =

    1.0000
    1.0000
    0.8333


recall =

    1.0000
    0.8000
    1.0000
```

This screenshot is about the confusion matrix CM. It classify the class into true positive, true negative, false positive, false negative. It also shows the predicted labels are from positive class and its accuracy.

C. NSR of two E-Commerce site using hybrid dictionary and SVM Classifier



In above screenshot, it is shown that for two combine ecommerce site i.e. amazon and flip kart is selected and selecting hybrid dictionary using ANN classifier method the above figure shows that amazon SR is 0.76 and flip kart NSR is 0.86 and ANN accuracy 96.57% it takes more time taken 18.21 seconds than SVM classifier and both class belong to positive thus words counts and analysis has been shown in above figure using mat lab.

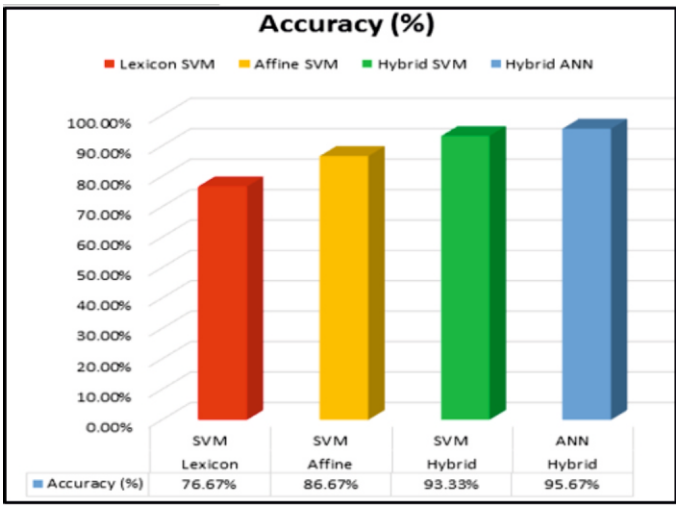
D. Table of Result analysis

Dictionary	Classifier	Time (Sec)	Accuracy (%)
Lexicon	SVM	3.25	76.67%
Affine	SVM	3.48	86.67%
Hybrid	SVM	3.21	93.33%
Hybrid	ANN	18.21	96.57%

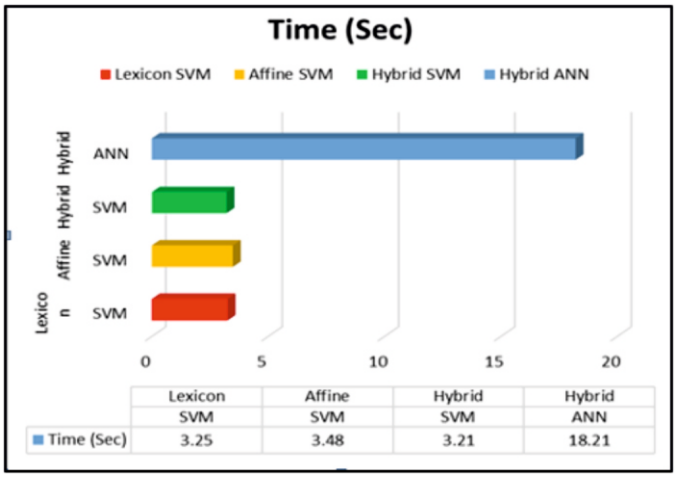
E. Accuracy and Time Graph

In below screenshot, it is shown that accuracy of SVM with Lexicon, Affin and Hybrid dictionary ANN with Hybrid. it is also shown that SVM with Hybrid gives better performance as compared to both lexicon and affin dictionary. Thus ANN also gives good accuracy but it takes more time for process while SVM takes less time to process and gives better performance 93.33%.

Accuracy Graph



Time Graph



In this screenshot it is shown that SVM takes less time while ANN takes more time to process.

CONCLUSION

In this research, dictionary concept for sentiment score calculation is used and it has shown that by using external dictionaries (Affine, Lexicon and Hybrid) of words which has polarity score to capture the sentiment "polarity in everyday tweets. It is concluded that by using hybrid dictionary which contains more words with polarity score than lexicon and affine dictionary and it also contains emoticon feature with score and SVM classifier gives better accuracy 93.33% and it takes less time to process while using ANN classifier it takes more time to process due to iterations. Experiment results show hybrid dictionary with SVM classifier gives better performance in less time.

REFERENCES

[1] Lokmanyathilak Govindan Shankar Selvan, Tang-Sheng Moh, "A Framework for Fast-Feedback Opinion Mining on Twitter Data Streams", IEEE 2015

[2] Gökhan Solahtarolu,"Analysis and Prediction of E-Customers' Behavior by Mining Clickstream Data", IEEE 2015

[3] Koith Douglas Stuart and Macioj "Intelligent Opinion Mining and Sentiment Analysis Using ANN", Springer 2015

[4] G. Vinodhini, R M Chandrasekaran "Opinion mining using principal component analysis based ensemble model for e-commerce application", Springer 2014

[5] Divakar Yadav, Geetika Gautam "Sentiment Analysis of Twitter Data Using ML Approaches and Semantic Analysis", IEEE 2014

[6] Jun Yang, Lan Jiang, ChongJun Wan and Junyuan Xie "Multi-Label Emotion Classification for Tweets in Weibo: Chinese site", IEEE 2014

[7] Hauma Isah,Paul Trundle,Daneil Neagu "Social media analytics for product safety Using text mining and sentiment analysis", IEEE 2014

[8] Neethu M S, Rajasree R "Sentiment Analysis in Twitter using Machine Learning Techniques", IEEE 2013.

[9] Xujuan Zhou, Xiahui Tao, Jianming, Zhemyu "Sentiment Analysis on Tweets for Social Events", IEEE 2013

[10] Uma Nagarsekar, Priyanka Kulkarni "Emotion Detection from "The SMS of the Internet", IEEE 2013

[11] K Unnumalia "Analysis of product using web", Elsevier 2012

[12] M. Rushdi Saleh, M.T. Martín-Valdivia, A. Montejo-Ráez, L.A. Ureña-López "Experiments with SVM to classify opinions in different domains" ,Elsevier 2011

[13] Ziqiong Zhang, Qiang Ye, Zili Zhang, Yijun Li "Sentiment classification of Internet restaurant reviews written in Cantonese" Elsevier 2011

[11] A Survey Paper on "Social Media E-Customers Behaviour Mining", IJSART- Volume-2 Issue-12, Dec-2016, ISSN: 2395-1052

[12] M. Rushdi Saleh, M.T. Martín-Valdivia, A. Montejo-Ráez, L.A. Ureña-López "Experiments with SVM to classify opinions in different domains" ,Elsevier 2011

[13] Ziqiong Zhang, Qiang Ye, Zili Zhang, Yijun Li "Sentiment classification of Internet restaurant reviews written in Cantonese" Elsevier 2011